Submitted to the Annals of Applied Statistics
arXiv: arXiv:0000.0000

DIAGNOSIS-GROUP-SPECIFIC TRANSITIONAL CARE PROGRAM RECOMMENDATIONS FOR THIRTY-DAY REHOSPITALIZATION REDUCTION

By Menggang Yu*, Chensheng Kuang* Jared D. Huling† and Maureen Smith*

*University of Wisconsin-Madison** *and University of Minnesota*†

Thirty-day rehospitalization rate is a well-studied and important measure reflecting the overall performance of health systems. Recently, transitional care (TC) programs have been initiated to reduce avoidable rehospitalizations. These programs typically ask nurses to follow-up with patients after the hospitalization to manage issues and reduce the risk of rehospitalizations during health care transitions. As rehospitalization is a complex process that depends on many factors, it is unlikely that these interventions are effective for all patients across a diverse population. In this paper, we consider individualized intervention or treatment recommendation rules (ITRs) aimed at maximizing overall treatment effectiveness. We investigate our approach in a setting where patients are divided into two diagnosis related groups, medically complicated and uncomplicated. As the treatment effects can greatly vary between the two groups, we allow our recommendation rules to be group specific. In particular, our approach can accommodate scale differences in treatment effects and utilize a tuning parameter to drive the similarity of the estimated ITRs between groups. Computation is achieved by transforming our problem into a form solvable by existing software and a wrapper R package is developed for our proposed treatment recommendation framework. We conduct extensive evaluation through both simulation studies and analysis of a TC program.

1. Introduction. Early hospital readmission has been recognized as a common and costly occurrence, particularly among elderly and high-risk patients. One in five Medicare beneficiaries is readmitted within 30 days at a cost of more than $26 billion per year (Betancourt, Tan-McGrory and Kenst, 2015; Jencks, Williams and Coleman, 2009), with *avoidable* readmissions estimated to cost as much as $17 billion per year (Rau, 2014). To encourage improvement in the quality of care and a reduction in unnecessary health expense, policymakers, reimbursement strategists, and the US government

Keywords and phrases: Heterogeneity of treatment effect, Observational data, Rehospitalization, Subgroup identification, Data integration

have thus made reducing 30-day hospital readmissions a national priority. Through Congressional direction and executive initiatives, Medicare has begun implementing incentives to reduce hospital readmissions. One example is the Hospital Readmission Reduction Program (HRRP) (McIlvennan, Eapen and Allen, 2015), which financially penalizes hospitals with relatively high rates of Medicare readmissions. Another is the opportunity for participation in the Medicare Shared Savings Program as an incentive.

Consequently, hospitals and health systems have been focusing on reducing avoidable readmissions (Bradley et al., 2012; Donzé et al., 2013; Bradley et al., 2013; Cloonan, Wood and Riley, 2013; Leppin et al., 2014; Kripalani et al., 2014; Stevens, 2015), including the development of new interventions targeted towards reducing 30-day readmissions (Hansen et al., 2011). A key aspect contributing to high readmission rates is the lack of a 'continuum of care' (Evashwick, 2005; Fox et al., 2000; Norrving and Kissela, 2013). Transitional care (TC) interventions are a behavioral medicine approach to filling this gap. TC encompasses a broad range of services and environments designed to promote the safe and timely passage of patients between levels of health care and across different care settings (Naylor et al., 2011). High-quality TC is especially important for older adults with multiple chronic conditions and complex therapeutic regimens, as well as for their family caregivers. These patients typically receive care from many providers and move frequently within health care settings. A growing body of evidence suggests that they are particularly vulnerable to breakdowns in care and thus have the greatest need for TC services. Poor "handoff" of these older adults and their family caregivers from hospital to home has been linked to adverse events, low satisfaction with care, and high rehospitalization rates. TC programs typically ask a health worker (e.g., nurse) to follow-up with patients by telephone or in-person after a hospitalization to manage issues and reduce the risk of rehospitalizations during health care transitions (Naylor et al., 2011; Coleman et al., 2006; Kind et al., 2016).

Emerging evidence from the literature indicates that hospital readmission is a complicated process that can depend on many factors influenced not only by hospital environment, but also by variables outside of hospitals' direct control such as policy environment, social determinants, and patient lifestyle. This implies that any single program is not likely to be a silver bullet for solving the readmissions problem. In fact, it is expected that many health system interventions do not work well for everyone. This, combined with the reality that health systems often do not have adequate resources to enroll an entire population make it imperative for health systems to identify *which* patients are likely to benefit from an intervention or treatment. Thus, in this

work we focus on the identification of which patients are likely to benefit from the TC program. In particular, we aim to do this by estimating optimal individualized treatment rules (ITRs), which map patient characteristics to enrollment decisions in a manner that optimizes overall patient outcomes. The ITRs could be used to inform enrollment recommendations for patients. By better targeting patients for enrollment, we aim to optimize the program's efficiency and effectiveness, thus achieving better health outcomes for the population as well as cost savings for the hospital.

A major issue in the analysis of TC is that there are intrinsic differences in inpatient risk that complicate the effectiveness of the program. During the stay of a patient, a Diagnosis-Related Group (DRG) is determined and assigned to each patient. Based on the assigned DRG, patients can be split into medically uncomplicated and medically complicated groups; this DRG categorization scheme will be available at www.hipxchange.org/DRG. TC itself involves different administrative steps for patients in medically complicated DRGs than for medically uncomplicated DRGs. Figure 1 illustrates the 30-day readmission rates for the medically uncomplicated and medically complicated groups of patients. In both groups, the TC patients had lower readmission rates compared with their counterparts in the control arm, however the reduction depends strongly on the groups. Further, from the raw covariate values in the unweighted column in Table 1, the comorbidity profiles, baseline hospitalizations, baseline health care payments, and demographic information are all highly different between the two groups.

Given that the intervention itself is different for medically complicated versus medically uncomplicated patients and given the significant heterogeneity between the two groups of patients themselves, these differences should be taken into account when constructing ITRs for TC. Although there are major differences between medically complicated and uncomplicated patients, the two groups are still under the purview of the same health system and thus share influence from system level and geographic factors. Ignoring these similarities may lead to a loss of efficiency. Our aim is thus to construct an ITR estimation procedure that naturally accommodates these group differences while allowing for similarity in the ITRs between groups. Due to these intrinsic differences and the multi-faceted and personal nature of TC, the intervention itself is expected to be implemented differently for medically complicated versus uncomplicated patients. This complex structure necessitates careful handling to ensure interpretability and coherence of an analysis in a manner that allows for differences in treatment for different groups.

A rich variety of methods have recently been developed to derive individualized treatment or intervention rules based on patient characteristics.

	Unweighted Summary						Weighted Summary					
	Complicated			Uncomplicated			Complicated			Uncomplicated		
Variable	TC	Ctrl	p	TC	Ctrl	p	TC	Ctrl	p	TC	Ctrl	p
Age	78.14	74.87	< .001	78.82	76.37	< .001	76.83	76.83	.992	77.48	77.57	.760
Sex (female)	0.57	0.49	.018	0.61	0.54	< .001	0.52	0.53	.793	0.56	0.57	.591
Race (white)	0.91	0.93	.213	0.95	0.95	.945	0.92	0.92	.726	0.95	0.95	.952
Base Hosp	0.71	1.13	< .001	0.45	0.69	< .001	0.91	0.85	.547	0.52	0.51	.644
Base Payments	17.87	26.61	< .001	10.20	14.36	< .001	22.23	20.83	.542	11.28	11.00	.740
Base ED Visits	0.86	1.07	.076	0.83	0.94	.038	0.95	0.96	.948	0.89	0.87	.711
CHF	0.35	0.28	.023	0.24	0.18	< .001	0.36	0.34	.549	0.20	0.20	.962
COPD	0.36	0.38	.576	0.32	0.28	.014	0.36	0.37	.793	0.29	0.29	.788
CKD	0.51	0.52	1.00	0.37	0.35	.297	0.53	0.52	.845	0.35	0.35	.850
ESRD	0.09	0.14	.038	0.02	0.04	< .001	0.12	0.12	.801	0.03	0.02	.605
Diab(comp)	0.08	0.10	.510	0.09	0.10	.510	0.08	0.08	.878	0.10	0.10	.961
Diab(nocomp)	0.26	0.19	.010	0.14	0.13	.498	0.22	0.24	.653	0.14	0.14	.946
Depression	0.21	0.20	.737	0.20	0.21	.816	0.20	0.20	.833	0.21	0.21	.522
Obesity	0.20	0.16	.165	0.14	0.15	.266	0.18	0.18	.828	0.15	0.14	.587

TABLE 1

Baseline characteristics of medical complicated and uncomplicated patients in TC and control groups. The left hand side, labeled as the unweighted analysis, indicates the raw numbers before weighting by propensity scores and the right hand side, labeled as the weighted analysis, indicates weighted averages based on the inverse of the propensity score. P-values indicate differences in covariate values and are calculated using (weighted) t-tests for continuous covariates and (weighted) chi-square tests for discrete covariates.

Methods for estimation of ITRs roughly fall into one of two categories: i) methods which work by building an outcome model including both covariate main effects and interaction terms between covariates and treatment and derive the ITR by inverting the outcome model (Kehl and Ulm, 2006; Imai and Ratkovic, 2013; Qian and Murphy, 2011) and ii) methods which bypass outcome modeling and estimate the ITR directly (Lipkovich, Dmitrienko and B D'Agostino Sr, 2017; Zhao et al., 2012; Zhang et al., 2012; Chen et al., 2017). Under the latter approach, Zhang et al. (2012) and Zhao et al. (2012) proposed to estimate the ITR via a classification framework which sidesteps the direct modeling of the outcome, thus mitigating bias arising from misspecified outcome models. Even though many subgroup identification methods have been proved to be effective, there is little research directly applicable to the setting of the TC problem where potentially different ITRs are required for different groups of patients. Regarding heterogeneity of ITRs themselves, Shi et al. (2018) utilized an approach that estimates ITRs in a way that accounts for differences between fundamentally different groups of subjects, however it results in a universal or group-invariant ITR that applies to all groups.

To account for fundamentally different groups of patients, one could consider applying an existing ITR estimation approach to each patient group and obtain the corresponding group-specific ITR. However, the main concern

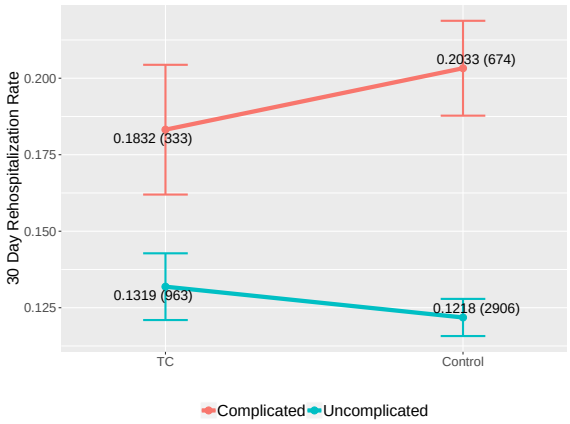

Fig 1: Displayed are the average unadjusted outcomes within the intervention groups stratified by medically complicated versus medically uncomplicated. Confidence intervals are at the 95% level.

of such an approach is that it may lack sufficient power, as the discovery of interactions between treatment and covariates often requires large sample sizes. Consequently, the ITR of a small group might be poorly estimated. To address this concern, we propose to simultaneously estimate different ITRs based on the assumption that some sort of commonality exists between the two (or more) groups. For example, group-structured penalties can be used to exploit similarities in variable selection among ITRs so as to boost estimation power, and effect estimates across groups can be shrunk to be more similar to each other when warranted. Our proposed approach allows the ITR for each patient group to be potentially different. In particular, our approach utilizes a tuning parameter which drives how similar the estimated ITRs are between groups and its value is chosen in a data-driven manner. Further, since the outcomes and covariates may be on entirely different scales between medically complicated and uncomplicated groups, we adopt the framework of Zhang et al. (2012), as it provides a natural mechanism for adjusting for scaling differences. Our framework is also applicable more broadly to different scenarios, such as ITR estimation for meta-analysis or the analysis of multiple outcomes.

The rest of the paper is organized as follows. In Section 2, we first review a weighted classification framework for ITR estimation, derive an extension of it that allows for efficient estimation, and finally extend it to the motivating setting of multiple patient groups. In Section 3, we introduce a reparameterization of our model that enables efficient computation via existing software.

In Section 4, we assess the performance of our method via simulation studies. In Section 5, the proposed method is utilized in an analysis of a TC program. Finally, we conclude in Section 6 with a summary and discussion of potentially useful extensions.

2. Methodology. In this section, we first review the weighted contrast classification framework introduced by Zhang et al. (2012) and develop an extension of it that allows for multiple patient groups. In general, the weighted classification framework of Zhang et al. (2012), which includes the well-known outcome weighted learning method (Zhao et al., 2012) as a special case, transforms the problem of estimating an ITR into a classification problem that bypasses the need to model the full outcome regression relationship. We use the potential outcome notations of the Rubin Causal Model (Rubin, 2005; Holland, 1986) in our development.

2.1. *Weighted contrast classification.* Consider a study with n subjects receiving either a treatment or a control. Let A be the treatment indicator, with $A = 0$ and $A = 1$ indicating control and treatment respectively. For each subject, we observe a p-dimensional row vector of baseline covariates $\boldsymbol{X}$ and an outcome Y. Let $Y^{(0)}$ and $Y^{(1)}$ be the potential outcomes corresponding to $A = 0$ and $A = 1$ respectively. Under the Stable Unit Treatment Value Assumption (Cox, 1958), the observed outcome Y can then be written as $Y = Y^{(1)}A + Y^{(0)}(1 - A)$. A treatment rule g is a function that maps from the space of $\boldsymbol{X}$ to $\{0, 1\}$. Given a subject with $\boldsymbol{X} = \boldsymbol{x}$, under the treatment rule g, he/she is recommended to receive $A = 1$ if $g(\boldsymbol{x}) = 1$ and $A = 0$ if $g(\boldsymbol{x}) = 0$. Then under g, we can express the observed outcome for a subject with baseline covariates $\boldsymbol{X}$ as $Y^{(g)}(\boldsymbol{X}) = Y^{(1)}g(\boldsymbol{X})+Y^{(0)}\{1-g(\boldsymbol{X})\}$. Without loss of generality, we assume that a larger value of the outcome is more favorable. Our goal is then to find $g^{\text{opt}} = \underset{g}{\text{argmax}} E\{Y^{(g)}(\boldsymbol{X})\}$.

In this paper, we also assume that there are no unmeasured confounders; i.e., $A \perp (Y^{(0)}, Y^{(1)})|\boldsymbol{X}$. Under this assumption, as was shown in Zhang et al. (2012), it is straightforward to deduce that

$$\begin{aligned} E\{Y^{(g)}(\boldsymbol{X})\} &= E\big[\, E(Y|A=1,\boldsymbol{X})g(\boldsymbol{X}) + E(Y|A=0,\boldsymbol{X})\{1-g(\boldsymbol{X})\}\,\big] \\ &= E[g(\boldsymbol{X})C(\boldsymbol{X}) + E(Y|A=0,\boldsymbol{X})] \end{aligned} \tag{2.1}$$

where

$$C(\boldsymbol{X}) \equiv E(Y|A=1,\boldsymbol{X}) - E(Y|A=0,\boldsymbol{X}) \tag{2.2}$$

is called the *contrast function* which reflects the expected treatment effect difference between $A = 0$ and $A = 1$. From (2.1), we see that finding g^{opt} is equivalent to finding the maximizer of $E[g(\boldsymbol{X})C(\boldsymbol{X})]$, i.e., $g^{\text{opt}} = \underset{g}{\text{argmax}} E\{g(\boldsymbol{X})C(\boldsymbol{X})\}$. Zhang et al. (2012) showed that

$$g(\boldsymbol{X})C(\boldsymbol{X}) = -|C(\boldsymbol{X})|\big[\,\mathbb{1}\{C(\boldsymbol{X}) > 0\} - g(\boldsymbol{X})\,\big]^2 + |C(\boldsymbol{X})|\mathbb{1}\{C(\boldsymbol{X}) > 0\},$$

where $\mathbb{1}\{\}$ is the indicator function. Thus, the original problem of finding the optimal treatment regime is transformed into finding the optimal classifier in a weighted classification problem; i.e.,

$$g^{\text{opt}} = \underset{g\in\mathcal{G}}{\text{argmin}} E\big(|C(\boldsymbol{X})|\,[\,\mathbb{1}\{C(\boldsymbol{X}) > 0\} - g(\boldsymbol{X})\,]^2\big) \tag{2.3}$$

where $\mathcal{G}$ is a collection of functions that take values in $\{0, 1\}$ and $g(\boldsymbol{X})$ is used to classify whether $C(\boldsymbol{X})$ is positive with $|C(\boldsymbol{X})|$ serving as weights. In general, the squared error loss $[\mathbb{1}\{C(\boldsymbol{X}) > 0\} - g(\boldsymbol{X})\,]^2$ could be replaced with an arbitrary loss function $\ell(\mathbb{1}\{C(\boldsymbol{X}) > 0\}, g(\boldsymbol{X}))$ which aims to compare how well $g(\boldsymbol{X})$ recovers the optimal decision rule $\mathbb{1}\{C(\boldsymbol{X}) > 0\}$. For example, one could utilize a logistic loss. We have found that the squared error loss works quite well in practice, perhaps due to its alignment with the form of (2.3). In particular, we have found in the context of the group-specific estimation context of Section 2.2 below that the squared error loss performs consistently better than the logistic loss and thus we use it for the remainder of this paper.

One major challenge in the optimization problem (2.3) is that it is computationally difficult to work with a function that only takes two values, 0 and 1, since the resulting optimization problem is neither continuous nor convex in covariates. We instead start from a function $\tilde{g}$ which comes from a family of predictors $\tilde{\mathcal{G}}$ taking values not restricted to only $\{0, 1\}$. For example, let $\tilde{g}^{\text{opt}}$ come from a linear predictor family $\tilde{\mathcal{G}} = \{\beta_0 + \boldsymbol{X}\boldsymbol{\beta} | \beta_0 \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p\}$. The treatment assignment then depends on whether $\tilde{g}(\boldsymbol{X}) - 1/2$ is greater than 0 or not, i.e., $g(\boldsymbol{X}) \equiv \mathbb{1}\{\tilde{g}(\boldsymbol{X}) - 1/2 > 0\}$ is the corresponding optimal treatment rule. Because the contrast function $C(\boldsymbol{X})$ is also unknown in practice, it must be replaced with an estimate $\widehat{C}(\boldsymbol{X}_i, Y_i, A_i)$. The resulting optimization problem:

$$\tilde{g}^{\text{opt}} = \underset{\tilde{g}\in\tilde{\mathcal{G}}}{\text{argmin}} \frac{1}{n}\sum_{i=1}^{n} |\widehat{C}(\boldsymbol{X}_i, Y_i, A_i)|\,\big[\mathbb{1}\{\widehat{C}(\boldsymbol{X}_i, Y_i, A_i) > 0\} - \tilde{g}(\boldsymbol{X}_i)\big]^2 \tag{2.4}$$

where $\widehat{C}(\boldsymbol{X}, Y, A)$ is an estimator of $C(\boldsymbol{X})$. Theorem 2.1 justifies the consistency of $\tilde{g}^{\text{opt}}$, hence of $\hat{g}^{\text{opt}}(\boldsymbol{X}) = \mathbb{1}\{\tilde{g}^{\text{opt}}(\boldsymbol{X}) - 1/2 > 0\}$, provided some extra conditions on $\widehat{C}(\boldsymbol{X}, Y, A)$ and $\tilde{\mathcal{G}}$ hold.

THEOREM 2.1. *Let*

$$g^{\infty}(\boldsymbol{X}) = \frac{E\{|\widehat{C}(\boldsymbol{X},Y,A)|\mathbb{1}\{\widehat{C}(\boldsymbol{X},Y,A)>0\}\big|\boldsymbol{X}\}}{E\big\{|\widehat{C}(\boldsymbol{X},Y,A)|\big|\boldsymbol{X}\big\}}.$$

If $g^{\infty}(\boldsymbol{X}) \in \tilde{\mathcal{G}}$ and $\tilde{\mathcal{G}}$ is a finite dimensional parametric family and $\widehat{C}(\boldsymbol{X},Y,A)$ is not 0 for all Y and A, then $\tilde{g}^{opt}$ in (2.4) converges to $g^{\infty}(\boldsymbol{X})$ as $n \to \infty$. Furthermore, if $\widehat{C}(\boldsymbol{X},Y,A)$ is an unbiased estimator of $C(\boldsymbol{X})$ and the solution of (2.3) is unique,

$$\tilde{g}^{opt}(\boldsymbol{X}) > 1/2 \Longleftrightarrow C(\boldsymbol{X}) > 0.$$

Therefore $\hat{g}^{opt}(\boldsymbol{X})$ is consistent for $C(\boldsymbol{X}) > 0$.

Proof of Theorem 2.1 is given in the Supplementary Material. We note that the assumption that $\widehat{C}(\boldsymbol{X},Y,A)$ is not 0 for all Y and A is quite mild and can easily be checked numerically. Clearly, if $\widehat{C}(\boldsymbol{X},Y,A)$ is an unbiased estimator of $C(\boldsymbol{X})$, then $\tilde{g}^{\text{opt}}(\boldsymbol{X}) > 1/2$ can be used to recover the sign of $C(\boldsymbol{X})$. Clearly, as long as the squared error loss is not flat, the solution of (2.3) is unique and this assumption required in Theorem 2.1 will hold. A widely used estimator for the contrast function $C(\boldsymbol{X})$ is the inverse probability weighted estimator (IPWE), which can be written as

$$\widehat{C}_{\text{IPWE}}(\boldsymbol{X},Y,A) = \frac{AY}{\pi(\boldsymbol{X})} - \frac{(1-A)Y}{1-\pi(\boldsymbol{X})} \tag{2.5}$$

where $\pi(\boldsymbol{X}) = P(A=1 \mid \boldsymbol{X})$ is the propensity score for a subject with baseline covariates $\boldsymbol{X}$. In a randomized trial, $\pi(\boldsymbol{X})$ is known and usually a constant. In an observational study, typically we estimate $\pi(\boldsymbol{X})$ using some binary outcome model such as logistic regression. It is not hard to show $\widehat{C}_{\text{IPWE}}$ is an unbiased estimator of $C(\boldsymbol{X})$ conditional on $\boldsymbol{X}$, i.e., $E\{\widehat{C}_{\text{IPWE}}(\boldsymbol{X},Y,A) \mid \boldsymbol{X}\} = C(\boldsymbol{X})$.

We note that from the definition of $C(\boldsymbol{X})$ in (2.2) that for any function $a(\boldsymbol{X})$ of $\boldsymbol{X}$,

$$C(\boldsymbol{X}) = E\{Y - a(\boldsymbol{X})|A=1,\boldsymbol{X}\} - E\{Y - a(\boldsymbol{X})|A=0,\boldsymbol{X}\}$$

Hence, we can also use the following outcome-shifted estimator for $C(\boldsymbol{X})$

$$\widehat{C}^{a}_{\text{IPWE}}(\boldsymbol{X},Y,A) = \frac{A[Y-a(\boldsymbol{X})]}{\pi(\boldsymbol{X})} - \frac{(1-A)[Y-a(\boldsymbol{X})]}{1-\pi(\boldsymbol{X})}, \tag{2.6}$$

which is also unbiased for $C(\boldsymbol{X})$. It is natural, then, to choose the function $a(\boldsymbol{X})$ which minimizes the conditional variance of $\widehat{C}^{a}_{\text{IPWE}}$. Proposition 2.2 provides the form of the optimal $a(\boldsymbol{X})$ whose proof is given in the Supplementary Material.

PROPOSITION 2.2. *For any* $a(\boldsymbol{X})$, $E\{\widehat{C}^a_{IPWE}(\boldsymbol{X}, Y, A) \,|\, \boldsymbol{X}\} = C(\boldsymbol{X})$. *Further denote* $a^{opt}(\boldsymbol{X}) = \arg\min_a Var\{\widehat{C}^a_{IPWE}(\boldsymbol{X}, Y, A) \,|\, \boldsymbol{X}\}$. *Then*

$$a^{opt}(\boldsymbol{X}) = \{1 - \pi(\boldsymbol{X})\}E(Y|A = 1, \boldsymbol{X}) + \pi(\boldsymbol{X})E(Y|A = 0, \boldsymbol{X}).$$

If $E(Y|A, \boldsymbol{X})$ is estimated by some statistical model $\hat{\mu}(\boldsymbol{X}, A)$ and $\pi(\boldsymbol{X})$ is estimated by $\hat{\pi}(\boldsymbol{X})$, then the resulting $\widehat{C}^a_{\text{IPWE}}$ from Proposition 2.2 is precisely the same as the augmented inverse probability weighted estimator (AIPWE) of Robins, Rotnitzky and Zhao (1994) and Zhang et al. (2012):

(2.7)
$$\widehat{C}_{\text{AIPWE}}(\boldsymbol{X}, Y, A) = \frac{AY}{\hat{\pi}(\boldsymbol{X})} - \frac{(1-A)Y}{1-\hat{\pi}(\boldsymbol{X})} - \frac{A - \hat{\pi}(\boldsymbol{X})}{\hat{\pi}(\boldsymbol{X})}\hat{\mu}(\boldsymbol{X}, 1) - \frac{A - \hat{\pi}(\boldsymbol{X})}{1-\hat{\pi}(\boldsymbol{X})}\hat{\mu}(\boldsymbol{X}, 0).$$

Although such an estimator of the contrast has appeared in the literature, our derivations show that it can be arrived at via a different perspective, i.e. variance reduction through outcome shifting. Due to this connection, it is clear that a^{opt} results in doubly robust estimation of the contrast. That is, as long as either the outcome model $\hat{\mu}(\boldsymbol{X}, A)$ or the propensity model $\hat{\pi}(\boldsymbol{X})$ is correctly estimated, $\widehat{C}_{\text{AIPWE}}$ will be a consistent estimator for $C(\boldsymbol{X})$. In addition, compared with $\widehat{C}_{\text{IPWE}}$, $\widehat{C}_{\text{AIPWE}}$ typically has smaller variance if $\hat{\pi}(\boldsymbol{X})$ is modeled correctly. In the remainder of the paper, we use $\widehat{C}_{\text{AIPWE}}$ to estimate $C(\boldsymbol{X})$.

2.2. *Extension to multiple patient groups.* We now turn our focus to the setting of multiple patient groups such as the TC scenario and propose an extension of the above framework that allows for simultaneous handling of high dimensional data while borrowing strength in estimation across patient groups. We use subscripts to denote subjects and superscripts to denote groups. That is, $\boldsymbol{X}_i^j$, Y_i^j, A_i^j denote the baseline covariates, the outcome, and the treatment indicator of the ith subject in the jth group. Further, let n_j denote the number of patients in the jth group and let $n = \sum_{j=1}^q n_j$ be the total number of patients, where q denotes the number of patient groups. In our TC scenario, $q = 2$.

One approach to handling group differences would be to apply the weighted contrast classification framework (2.4) for each patient group and estimate corresponding treatment rules. However, this approach may result in high variance due to smaller sample sizes in groups. Another approach would be to include a group indicator as a new covariate and consider its interactions with all other covariates. This approach is described mathematically in Section 4.2. However, this would ignore differential treatment assignment mechanisms by group or potential differences in treatment by group as are

both the case in our TC study. Thus, we propose a framework that jointly utilizes all available information while still allowing for group differences.

We propose to estimate the treatment rules $g^1, \ldots, g^q$ through the following optimization problem,

$$\tilde{g}^{1,\text{opt}}, \ldots, \tilde{g}^{q,\text{opt}} = \underset{\tilde{g}^1,\ldots,\tilde{g}^q \in \tilde{\mathcal{G}}}{\text{argmin}} \Big\{ \sum_{j=1}^{q} \frac{1}{2W^j} \sum_{i=1}^{n_j} |\widehat{C}^j(\boldsymbol{X}_i^j, Y_i^j, A_i^j)| \times \left[\mathbb{1}\{\widehat{C}^j(\boldsymbol{X}_i^j, Y_i^j, A_i^j) > 0\} - \tilde{g}^j(\boldsymbol{X}_i^j) \right]^2 + h(\tilde{g}^1, \ldots, \tilde{g}^q) \Big\} \tag{2.8}$$

where $\widehat{C}^j$ is the corresponding contrast estimator, W^j a standardization weight, and $\tilde{g}^j$ the treatment rule, for the jth group, and h is a regularization term. We discuss the choices for h and W^j below. Throughout the paper, we focus on a linear family of estimators for the ITRs, i.e., $\tilde{\mathcal{G}} = \{\beta_0 + \boldsymbol{X}^T\boldsymbol{\beta} | \beta_0 \in \mathcal{R}, \boldsymbol{\beta} \in \mathcal{R}^p\}$. Therefore we write $\tilde{g}^j(\boldsymbol{X}) = \beta_0^j + \boldsymbol{X}^T\boldsymbol{\beta}^j$.

The specific form of h is crucial to our approach and is the mechanism by which we borrow strength across the different groups in estimation of the group-specific ITRs. The function h simultaneously facilitates *variable selection*, by setting some terms in $\tilde{g}^j$ to zero, and *variance reduction*, by encouraging terms in $\tilde{g}^1, \ldots, \tilde{g}^q$ to be similar when warranted. Further, with the idea that important variables are likely to be important for most groups together, h also promotes simultaneous selection and removal of variables across the groups.

To describe our proposed approach for borrowing strength via h, we begin by decomposing each element of $\boldsymbol{\beta}^j = (\beta_1^j, \ldots, \beta_p^j)^T$ as $\beta_k^j = \mu_k + \delta_k^j$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$ are the effects common across all q groups and $\boldsymbol{\delta}^j = (\delta_1^j, \ldots, \delta_p^j)$ are the group-specific effects. Under this parameterization, the lasso penalization of the δ_k^j terms, $|\delta_k^j| = |\beta_k^j - \mu_k|$, is equivalent to a fused lasso penalty that encourages the effects β_k^j of group j to be similar to the common effects μ_k. This decomposition utilizes no reference group and is overparameterized. However, Ollier and Viallon (2017) observed that when a lasso penalty is utilized, it yields nearly equivalent results as using the "optimal" reference group. We further denote $\boldsymbol{\delta}_k = (\delta_k^1, \ldots, \delta_k^q)^T$ and $\boldsymbol{\delta} = (\boldsymbol{\delta}_1^T, \ldots, \boldsymbol{\delta}_p^T)^T$. For any vector $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots)$, denote its ℓ_d norm $||\boldsymbol{\eta}||_d \equiv (\sum_j |\eta_j|^d)^{1/d}$. Further, for two vectors $\boldsymbol{\eta}$ and $\boldsymbol{\nu}$ of equal length, denote $\boldsymbol{\eta} \odot \boldsymbol{\nu}$ as the element-wise product of $\boldsymbol{\eta}$ and $\boldsymbol{\nu}$. Adopting the notation

of the sparse group lasso (Simon et al., 2013a), we define

$$h(\tilde{g}^1, \ldots, \tilde{g}^q) = (1-\alpha)\lambda_1\sqrt{q}\left\{\sum_{k=1}^{p}\|(\mu_k, \boldsymbol{\tau}\odot\boldsymbol{\delta}_k)\|_2\right\} \tag{2.9}$$

$$+\,\alpha\lambda_1\left\{\|\boldsymbol{\mu}\|_1 + \sum_{j=1}^{q}\tau_j\|\boldsymbol{\delta}^j\|_1\right\}, \tag{2.10}$$

where λ_1 is a penalty parameter that controls the overall strength of penalization, $\alpha \in [0,1]$ encapsulates a trade-off between the group penalties and lasso penalties, and the terms $\boldsymbol{\tau} \equiv (\tau_1, \ldots, \tau_q)$ allow for differential penalization of the group-specific coefficients $\boldsymbol{\delta}^j$.

We now describe the different tuning parameters in h and discuss their impact on the resulting ITRs. Both (2.9) and (2.10) aid in borrowing strength across groups while also performing variable selection. The terms τ_j and α control how much information is borrowed across the groups. Smaller values of α encourage variables to be selected or set to zero simultaneously across all groups and larger values of τ_j result in the effects $\boldsymbol{\beta}^j$ to be more similar to $\boldsymbol{\mu}$. On the other hand, larger values of α and smaller values of the τ_j terms result in less information being borrowed across groups. Thus if variables are either important in common or have similar effect sizes across groups, both (2.9) and (2.10) will aid in reducing variance in estimation of the effects. Following Ollier and Viallon (2017), we set each $\tau_j = \tau_0(n_j/n)^{1/2}$ for some $\tau_0 > 0$ so that the penalization is only influenced by the sample sizes of the groups. As per Simon et al. (2013a), is not recommended to choose α via cross validation, but rather the user should choose α based on expected levels of group sparsity. See Simon et al. (2013a) for extended discussion.

Using penalties (2.9) and (2.10) under the weighted classification framework is sensible since differential treatment effect sizes across different groups can be directly controlled. Because β_0^j and $\boldsymbol{\beta}^j$ are the coefficients in linear discriminant functions which classify between 0 and 1 as opposed to effect estimation in an outcome regression model, we can more reasonably assume that the magnitudes of the discriminant will be less variable across groups and thus penalties like (2.9) and (2.10) are likely to be acting on similar scales. This issue will be further illustrated through extensive simulation studies.

We note that a regularization structure similar to (2.9) and (2.10) can be used to improve estimation efficiency for the propensity scores and contrast functions via joint estimation across groups if parametric models are used. However, if non-parametric or otherwise flexible machine learning approaches

are used for estimation for the propensity scores and contrast functions, joint estimation may be of less benefit.

As we have seen in Figure 1, the contrast functions may not be of the same magnitude across groups due to population differences, e.g. the treatment may be much more effective in one patient group than in other groups. For this reason, we introduce the standardization weight W^j and one sensible choice is

$$W^j = \sum_{i=1}^{n_j} |\widehat{C}^j(\boldsymbol{X}_i^j, Y_i^j, A_i^j)| \tag{2.11}$$

which reflects the magnitude in both group sizes and contrast function values. Of course one can also choose $W^j = n_j$ with adjustment for the group size or $W^j \equiv 1$ without any adjustment.

2.3. *Universal ITR rule.* When a universal treatment rule is desired for all patient groups, we propose to solve the following optimization problem,

$$\tilde{g}^{\text{opt}} = \underset{\tilde{g}\in\tilde{\mathcal{G}}}{\text{argmin}} \Big\{ \sum_{j=1}^{q} \frac{1}{2W^j} \sum_{i=1}^{n_j} |\widehat{C}^j(\boldsymbol{X}_i^j, Y_i^j, A_i^j)| \big[\mathbb{1}\{\widehat{C}^j(\boldsymbol{X}_i^j, Y_i^j, A_i^j) > 0\} - \tilde{g}(\boldsymbol{X}_i^j) \big]^2 + h(\tilde{g}) \Big\} \tag{2.12}$$

where $\tilde{g}$ is the universal rule used for all groups. We denote $\tilde{g}(\boldsymbol{X}) = \beta_0 + \boldsymbol{X}^T\boldsymbol{\beta}$. W^j is defined in (2.11). Now the penalty h in (2.12) mainly aids in variable selection as there is no need for group regularization or penalties that encourage similarity of coefficients across groups. The estimator (2.12) still accounts for group-wise variability in treatment selection processes and outcomes, but results in a single ITR that works as well as possible across all groups.

A long list of options is available for h, such as the lasso penalty (Tibshirani, 1996), elastic net (Zou and Hastie, 2005), SCAD (Fan and Li, 2001), or MCP (Zhang, 2010), among many others. In particular, we choose the lasso in our numerical studies for comparison because of its established theoretical underpinnings and efficient computational algorithms. That is,

$$h = \lambda_{\text{universal}} \|\boldsymbol{\beta}\|_1 \tag{2.13}$$

where $\lambda_{\text{universal}}$ is a tuning parameter that guides the degree of penalization.

3. Computation. The major computational challenge in the proposed framework is to solve the optimization problem (2.8) when both (2.9) and (2.10) are included in the regularization term h. In this section, we will show how to utilize existing software to minimize (2.8) in this setting.

3.1. *Data preprocessing.* For notational simplicity, let $|\widehat{C}_i^j|$ denote $|\widehat{C}^j(\boldsymbol{X}_i^j, Y_i^j, A_i^j)|$ and S_i^j denote $\mathbb{1}\{\widehat{C}^j(\boldsymbol{X}_i^j, Y_i^j, A_i^j) > 0\}$. First, we simplify the problem by removing all the β_0^j since they are not involved in either (2.9) or (2.10). This can be achieved by centering all the baseline covariates and S_i^j with weights $|\widehat{C}_i^j|$ within each patient group. We can further simplify the problem by absorbing the weights $|\widehat{C}_i^j|$ and W^j by eventually working with the following transformed quantities, that is, if we denote

$$\check{\boldsymbol{X}}_i^j \equiv \sqrt{\frac{|\widehat{C}_i^j|}{W^j}}\left\{\boldsymbol{X}_i^j - \frac{\sum_{i=1}^{n_j}|\widehat{C}_i^j|\boldsymbol{X}_i^j}{\sum_{i=1}^{n_j}|\widehat{C}_i^j|}\right\} \quad \text{and} \quad \check{S}_i^j \equiv \sqrt{\frac{|\widehat{C}_i^j|}{W^j}}\left\{S_i^j - \frac{\sum_{i=1}^{n_j}|\widehat{C}_i^j|S_i^j}{\sum_{i=1}^{n_j}|\widehat{C}_i^j|}\right\},$$

then we solve the following problem

$$\min_{\boldsymbol{\beta^1},\ldots,\boldsymbol{\beta^q}} \left\{ \frac{1}{2}\sum_{j=1}^{q}\sum_{i=1}^{n_j}\left[\check{S}_i^j - \check{\boldsymbol{X}}_i^{jT}\boldsymbol{\beta^j}\right]^2 + (1-\alpha)\lambda_1\sqrt{q}\sum_{k=1}^{p}\sqrt{\mu_k^2 + (\tau_1\delta_k^1)^2 + \cdots + (\tau_q\delta_k^q)^2} \right.$$

(3.1)

$$\left. + \alpha\lambda_1\sum_{k=1}^{p}\left(|\mu_k| + \sum_{j=1}^{q}\tau_j|\delta_k^j|\right)\right\}.$$

3.2. *Computation via data transformation.* In the same vein as Ollier and Viallon (2017), we transform the design matrix so as to allow for computation using existing software for the sparse group lasso. Define the transformed design matrix as

$$\widetilde{\mathbf{X}} = \begin{pmatrix} \check{\mathbf{X}}^1 & \check{\mathbf{X}}^1/\tau_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \check{\mathbf{X}}^2 & \mathbf{0} & \check{\mathbf{X}}^2/\tau_2 & \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \check{\mathbf{X}}^q & \mathbf{0} & \ldots & \mathbf{0} & \check{\mathbf{X}}^q/\tau_q, \end{pmatrix}$$

where $\check{\mathbf{X}}^j$ is the design matrix for group j with row i as $\check{\boldsymbol{X}}_i^{jT}$. Further define $\widetilde{\mathbf{S}} = (\check{\mathbf{S}}^{1T}, \ldots, \check{\mathbf{S}}^{qT})^T$, where $\check{\mathbf{S}}^j = (\check{S}_1^j, \ldots, \check{S}_{n_j}^j)$. Then the target function (3.1) can be minimized by providing $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{S}}$ to the `SGL` function from the `SGL R` package (Simon et al., 2013b) or other software for the sparse group lasso.

To facilitate the usage of our method, we provide an `R` package "mpersonalized" that implements both the proposed framework in Section 2.2 and its single-intervention-rule variant in Section 2.3. "mpersonalized" package can be installed from the Github repository `https://github.com/chenshengkuang/mpersonalized`.

4. Simulation. In order to fully demonstrate the effectiveness of the proposed framework, we consider simulation scenarios where the true optimal ITRs in different patient groups are similar to each other in some sense. Hence, we include both (2.9) and (2.10) to boost estimation efficiency. The `SGL` package will be utilized for computation.

4.1. *Simulation Setup.* Even though ITRs only depend on the contrast functions, the outcome Y^j for patient group j was generated from models including main effects of covariates. Therefore, we use the following model to generate outcomes:

$$Y_i^j = \gamma_0^j + \boldsymbol{X}_i^{jT}\boldsymbol{\gamma}^j + A_i^j\left(\theta_0^j + \boldsymbol{X}_i^{jT}\boldsymbol{\theta}^j\right) + \epsilon_i^j \qquad i = 1, \ldots, n_j,\ j = 1, \ldots, q. \tag{4.1}$$

The values of γ_0^j, $\boldsymbol{\gamma}^j$, θ_0^j, $\boldsymbol{\theta}^j$ depend on the scenarios we will discuss below. Here we use different symbols, θ_0^j and $\boldsymbol{\theta}^j$, rather than β_0^j and $\boldsymbol{\beta}^j$, for the interaction coefficients because the focus is on recovering the sign of $\theta_0^j + \boldsymbol{X}_i^{jT}\boldsymbol{\theta}^j$ instead of estimation of θ_0^j and $\boldsymbol{\theta}^j$. The covariates $\boldsymbol{X}_i^j$ were generated independently from standard normal distributions. The treatment A_i^j was assigned with confounding based on $\boldsymbol{X}_i^j$, with $\text{logit}(\pi(\boldsymbol{X}_i^j)) = \boldsymbol{X}_i^{jT}\boldsymbol{\beta}_\pi$, where the first 8 elements of $\boldsymbol{\beta}_\pi$ were chosen uniformly at random from $\{-0.5, 0.5\}$ and the rest were 0. The "error" terms ϵ_i^j are generated from a Gamma distribution with shape s^j and rate r^j parameters that vary by group with $s^j \sim \text{Unif}(0.5, 4)$ and $r^j \sim \text{Unif}(0.1, 0.75)$. Here, $E(\epsilon_i^j) = s^j/r^j$, so the error terms' means act to change the intercept of the response. Both this and the difference in the distributions across the groups adds additional inter-group heterogeneity. As our motivating study involves a binary outcome, we present additional simulation studies with binary outcomes in the Supplementary Material. In the Supplementary Material, we additionally investigate the effect of the signal strength of the main effects on performance.

We assessed the performance of our method on simulated data in five scenarios, which differed in the number of covariates, group sizes, and treatment effects. In each scenario, we generated 200 datasets and included 6 different patient groups. In Scenarios 1 to 4, groups were balanced in size with each group containing $n_j = 100$ observations. In Scenario 5, group 1 had 500 observations while the other groups had 100 each. The number of baseline covariates p was 50 in Scenarios 1, 2, and 5, and 100 in Scenarios 3 and 4.

For the main effect coefficients, only γ_0^j and the first 11 elements of $\boldsymbol{\gamma}^j$ were nonzero. In all 5 scenarios, we generated the nonzero part of γ_0^j and $\boldsymbol{\gamma}^j$ in the following manner: first some common main effect coefficients γ_0 and

$\boldsymbol{\gamma}$ were generated, then for each group index j, γ_0^j and $\boldsymbol{\gamma}^j$ were generated by adding some perturbations to γ_0 and $\boldsymbol{\gamma}$. In this way we introduced both similarity and heterogeneity among the groups. The perturbations were generated randomly for each data set and took value of either 0.2 or -0.2 with equal probabilities. We show one realization of the nonzero parts of γ_0^j and $\boldsymbol{\gamma}^j$ in Table 2.

	γ_0^j	γ_1^j	γ_2^j	γ_3^j	γ_4^j	γ_5^j	γ_6^j	γ_7^j	γ_8^j	γ_9^j	γ_{10}^j	γ_{11}^j
Population	-2.0	-2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	-2.0	2.0
$j=1$	-2.2	-2.2	2.2	1.8	2.2	2.2	2.2	2.2	2.2	1.8	-2.2	1.8
$j=2$	-2.2	-1.8	1.8	1.8	2.2	1.8	2.2	2.2	1.8	2.2	-1.8	1.8
$j=3$	-2.2	-2.2	1.8	2.2	1.8	1.8	2.2	2.2	2.2	1.8	-2.2	2.2
$j=4$	-1.8	-2.2	1.8	1.8	2.2	2.2	1.8	1.8	1.8	1.8	-1.8	2.2
$j=5$	-1.8	-1.8	1.8	2.2	2.2	1.8	2.2	1.8	2.2	2.2	-2.2	2.2
$j=6$	-2.2	-2.2	2.2	2.2	1.8	2.2	2.2	1.8	2.2	2.2	-2.2	2.2

TABLE 2

Displayed is one realization of main effect coefficients (with γ_k^j denoting the kth element of $\boldsymbol{\gamma}^j$ for $k = 1, \ldots, 11$).

For the interaction effect coefficients, in Scenario 1, θ_0^j and $\boldsymbol{\theta}^j$ were generated in a similar way as γ_0^j and $\boldsymbol{\gamma}^j$. We first fixed some common interaction effect coefficients θ_0 and $\boldsymbol{\theta}$, and then added perturbations to generate θ_0^j and $\boldsymbol{\theta}^j$ in each simulated data set. Only θ_0^j and the first 5 elements of $\boldsymbol{\theta}^j$ were nonzero. The perturbations for the interaction effect coefficients was either 0.5 or -0.5. In Scenario 2, compared with Scenario 1, we added some group-specific effects. In Scenario 3, based on the coefficients of Scenario 2, we further increased the number of covariates to 100. In Scenario 4, θ_0^j and $\boldsymbol{\theta}^j$ were generated in the same manner as in Scenario 3 for $j = 2$ to 6, but θ_0^1 and $\boldsymbol{\theta}^1$ were multiplied by 4 instead. This scenario mimics the setting where a treatment has much stronger effect in one patient group than in other groups. In Scenario 5, we increased the group-specific treatment effects compared with Scenario 2 and the groups were unbalanced in sample sizes. The detailed group sizes and number of covariates for each scenario can be found in Table 1 of the Supplementary Material, which also includes one realization of the nonzero parts of θ_0^j and $\boldsymbol{\theta}^j$ for all five scenarios.

4.2. *Comparator Methods.* Both the proposed joint analysis framework with separate ITR rules for different patient groups in (2.8) and its universal or group-invariant ITR version in (2.12) were applied to the simulated datasets. For simplicity, they are denoted as "SITR.joint" and "UITR.joint" respectively. In addition, they are also compared to six other approaches.

(i) "SITR.naive": separate analysis via the contrast classification framework (Zhang et al., 2012; Zhao et al., 2012). The method estimates the ITR for each group separately through the contrast classification and applies a lasso penalty for variable selection. Specifically, it estimates β_0^j and $\boldsymbol{\beta}^j$ for all $j = 1, \ldots, q$ from

$$\min_{\beta_0^j, \boldsymbol{\beta}^j} \left\{ \sum_{i=1}^{n_j} |\widehat{C}^j(\boldsymbol{X}_i^j, Y_i^j, A_i^j)| \left[\mathbb{1}\{\widehat{C}^j(\boldsymbol{X}_i^j, Y_i^j, A_i^j) > 0\} - \beta_0^j - \boldsymbol{X}_i^{jT} \boldsymbol{\beta}^j \right]^2 + \lambda^j \|\boldsymbol{\beta}^j\|_1 \right\}.$$

(ii) "UITR.naive": pooled analysis via the contrast classification framework (Zhang et al., 2012; Zhao et al., 2012). Multiple patient groups are first combined into one and then *"SITR.naive"* is applied to this pooled patient group. In this method, $\hat{\beta}_0^j$ and $\hat{\boldsymbol{\beta}}^j$ from different groups are the same for $j = 1, \ldots, q$ and can be estimated from

$$\min_{\beta_0, \boldsymbol{\beta}} \left\{ \sum_{j=1}^{q} \sum_{i=1}^{n_j} |\widehat{C}^{\text{pool}}(\boldsymbol{X}_i^j, Y_i^j, A_i^j)| \left[\mathbb{1}\{\widehat{C}^{\text{pool}}(\boldsymbol{X}_i^j, Y_i^j, A_i^j) > 0\} - \beta_0 - \boldsymbol{X}_i^{jT} \boldsymbol{\beta} \right]^2 + \lambda^{\text{pool}} \|\boldsymbol{\beta}\|_1 \right\}.$$

Note that the contrast estimator $\widehat{C}^{\text{pool}}$ is also constructed over the pooled group.

(iii) "UITR.naive.ind": pooled analysis via the contrast classification framework (Zhang et al., 2012; Zhao et al., 2012). Multiple patient groups are first combined into one and then *"SITR.naive"* is applied to this pooled patient group and group indicators are included with their interactions with covariates. In this method, $\hat{\beta}_0^j$ and $\hat{\boldsymbol{\beta}}^j$ from different groups are not the same for $j = 1, \ldots, q$ due to the interactions with group indicators and can be estimated from

$$\min_{\beta_0, \boldsymbol{\beta}} \left\{ \sum_{j=1}^{q} \sum_{i=1}^{n_j} |\widehat{C}^{\text{pool}}(\boldsymbol{X}_i^j, Y_i^j, A_i^j)| \left[\mathbb{1}\{\widehat{C}^{\text{pool}}(\boldsymbol{X}_i^j, Y_i^j, A_i^j) > 0\} - \beta_0 - \widetilde{\boldsymbol{X}}_i^T \boldsymbol{\beta} \right]^2 + \lambda^{\text{pool}} \|\boldsymbol{\beta}\|_1 \right\},$$

where $\widetilde{\boldsymbol{X}}_i^T = (\boldsymbol{X}_i^T, I(G_i = 1), \ldots, I(G_i = q), I(G_i = 1)\boldsymbol{X}_i^T, \ldots, I(G_i = q)\boldsymbol{X}_i^T)$ contains all covariates and all interactions between covariates and group indicators. A key difference between this approach and "SITR.naive" is that the tuning parameter here is chosen uniformly across groups.

(iv) "SITR.reg": separate analysis via outcome regression modeling. For each patient group, two outcome models are fitted. One for the treatment arm and one for the control arm. Lasso is again used for variable selection and eventually ITR is determined by comparing the two mean

models. Specifically, for group j, the coefficients for the two outcome models are estimated from

$$\hat{\theta}^j_{0,A=0}, \hat{\boldsymbol{\theta}}^j_{A=0} = \min_{\theta^j_0, \boldsymbol{\theta}^j} \Big\{ \sum_{A^j_i=0} \big\{ Y^j_i - \theta^j_0 - \boldsymbol{X}^{j^T}_i \boldsymbol{\theta}^j \big\}^2 + \lambda^j_{A=0} \|\boldsymbol{\theta}^j\|_1 \Big\}$$

$$\hat{\theta}^j_{0,A=1}, \hat{\boldsymbol{\theta}}^j_{A=1} = \min_{\theta^j_0, \boldsymbol{\theta}^j} \Big\{ \sum_{A^j_i=1} \big\{ Y^j_i - \theta^j_0 - \boldsymbol{X}^{j^T}_i \boldsymbol{\theta}^j \big\}^2 + \lambda^j_{A=1} \|\boldsymbol{\theta}^j\|_1 \Big\}.$$

For a patient with baseline covariates $\boldsymbol{X}$, $A = 1$ is recommended if $\hat{\theta}^j_{0,A=1} + \boldsymbol{X}^T \hat{\boldsymbol{\theta}}^j_{A=1} - \hat{\theta}^j_{0,A=0} - \boldsymbol{X}^T \hat{\boldsymbol{\theta}}^j_{A=0} > 0$ and vice versa.

(v) "UITR.reg": pooled analysis via outcome regression modeling. Multiple groups are combined into one and then "SITR.reg" is applied. Similar to "UITR.naive", only one treatment rule is estimated for all groups. The coefficients of the outcome models are estimated from

$$\hat{\theta}_{0,A=0}, \hat{\boldsymbol{\theta}}_{A=0} = \min_{\theta_0, \boldsymbol{\theta}} \Big\{ \sum_{j=1}^{q} \sum_{A^j_i=0} \big\{ Y^j_i - \theta_0 - \boldsymbol{X}^{j^T}_i \boldsymbol{\theta} \big\}^2 + \lambda_{A=0} \|\boldsymbol{\theta}\|_1 \Big\}$$

$$\hat{\theta}_{0,A=1}, \hat{\boldsymbol{\theta}}^j_{A=1} = \min_{\theta_0, \boldsymbol{\theta}} \Big\{ \sum_{j=1}^{q} \sum_{A^j_i=1} \big\{ Y^j_i - \theta_0 - \boldsymbol{X}^{j^T}_i \boldsymbol{\theta} \big\}^2 + \lambda_{A=1} \|\boldsymbol{\theta}\|_1 \Big\}.$$

For a patient with baseline covariates $\boldsymbol{X}$, $A = 1$ is recommended if $\hat{\theta}_{0,A=1} + \boldsymbol{X}^T \hat{\boldsymbol{\theta}}_{A=1} - \hat{\theta}_{0,A=0} - \boldsymbol{X}^T \hat{\boldsymbol{\theta}}_{A=0} > 0$.

(vi) "UITR.reg.ind": pooled analysis via outcome regression modeling but with group by covariate interactions as with "UITR.naive.ind" so that the covariate vector used is $\widetilde{\boldsymbol{X}}_i$ as defined for "UITR.naive.ind".

4.3. *Implementation Details.* In each scenario, we generated validation datasets of the same sizes as training datasets for the purpose of tuning parameter selection in all methods. For "SITR.joint" and "UITR.joint", λ and τ_0 were chosen from a grid of numbers and the optimal pair minimized the following weighted loss in the validation dataset,

$$\sum_{j=1}^{q} \frac{1}{W^j} \sum_{i=1}^{n_j} |\widehat{C}^j(\boldsymbol{X}^j_i, Y^j_i, A^j_i)| \big[\mathbb{1}\{\widehat{C}^j(\boldsymbol{X}^j_i, Y^j_i, A^j_i) > 0\} - \hat{\beta}^j_0 - \boldsymbol{X}^{j^T}_i \hat{\boldsymbol{\beta}}^j \big]^2$$

where the contrast function $\widehat{C}^j(\boldsymbol{X}^j_i, Y^j_i, A^j_i)$ in the validation set was estimated independently of the training set. For "SITR.naive" and "UITR.naive", the tuning parameters were selected by minimizing the weighted losses in separate groups and in the pooled group of the validation set, respectively. For "SITR.reg" and "UITR.reg", the optimal penalty parameters minimized the mean square error of the outcome models in the validation dataset.

4.4. *Performance Evaluation.* For the contrast classification based methods, $\tilde{g}^j(\boldsymbol{X}) = \hat{\beta}_0^j + \boldsymbol{X}^T\hat{\boldsymbol{\beta}}^j$, and the estimated ITRs are $\hat{g}^j(\boldsymbol{X}) = \mathbb{1}\{\tilde{g}^j(\boldsymbol{X}) - 1/2 > 0\}$. On the other hand, for the outcome regression based methods, $\tilde{g}^j(\boldsymbol{X}) = \hat{\theta}_0^j + \boldsymbol{X}^T\hat{\boldsymbol{\theta}}^j$, and the estimated ITRs are $\hat{g}^j(\boldsymbol{X}) = \mathbb{1}\{\tilde{g}^j(\boldsymbol{X}) > 0\}$. This is because the former is based on classifiers for $\{0, 1\}$ whereas the latter is based on the interaction part $\theta_0 + \boldsymbol{X}^T\boldsymbol{\theta}$.

Results are evaluated using a large test set with $n_t = 10000$ subjects. The performance is compared through three measures. The first measure is the improvement in expected outcomes, or $E_{\boldsymbol{X}}\{Y^{(\hat{g})}(\boldsymbol{X})\} - E_{\boldsymbol{X}}\{Y^{(1)}(\boldsymbol{X})\}$. Let $\tilde{\boldsymbol{X}}_i$ be the baseline covariates of the ith subject in the test set. Then $E_{\boldsymbol{X}}\{Y^{(\hat{g})}(\boldsymbol{X})\} - E_{\boldsymbol{X}}\{Y^{(1)}(\boldsymbol{X})\}$ for group j is estimated by

$$\frac{1}{n_t}\sum_{i=1}^{n_t}\{\hat{g}^j(\tilde{\boldsymbol{X}}_i) - 1\}(\theta_0^j + \tilde{\boldsymbol{X}}_i{}^T\boldsymbol{\theta}^j),$$

where θ_0^j and $\boldsymbol{\theta}^j$ are the true parameters from the data generation model (4.1). The second measure is the correct treatment recommendation rate. Under the data generation model (4.1), it is basically the concordance between $\hat{g}^j(\tilde{\boldsymbol{X}}_i)$ and $\mathbb{1}\{\theta_0^j + \tilde{\boldsymbol{X}}_i{}^T\boldsymbol{\theta}^j > 0\}$ for $i = 1, \ldots, n_t$ and $j = 1, \ldots, q$. The third measure is rank correlation between estimated treatment effects and true treatment effects. This measure is estimated by the rank correlation between $\tilde{g}^j(\tilde{\boldsymbol{X}}_i)$ and $\theta_0^j + \tilde{\boldsymbol{X}}_i{}^T\boldsymbol{\theta}^j$.

4.5. *Simulation Results.* Figure 2 summarizes results under the three performance measures averaged over 6 groups based on the 200 simulations. Detailed study-specific tables of results and standard errors are available in the Supplementary Material. For the outcome improvement metric, the optimal results are displayed, which is the outcome improvement using the true underlying individual treatment effects. Note that for the concordance rate and rank correlation metrics, the optimal results are not displayed since they are simply 1 for both metrics. In general, "SITR.joint" outperforms the other methods in all scenarios, i.e. has the largest improvements in outcomes, highest concordance rates, and largest rank correlations with the true ITRs, except in Scenario 1 where "UITR.naive" and "UITR.reg" has comparable performance because the interaction effects are very similar across the groups. We also observe that "SITR.naive" and "SITR.reg" have the worst performance in all scenarios except in Scenario 5.

By comparing the results from Scenarios 2 and 3 which differ only in the number of useless covariates, we can also see that a larger $p = 100$ (Scenario 3) tends to worsen the performance of all the methods. In particular,

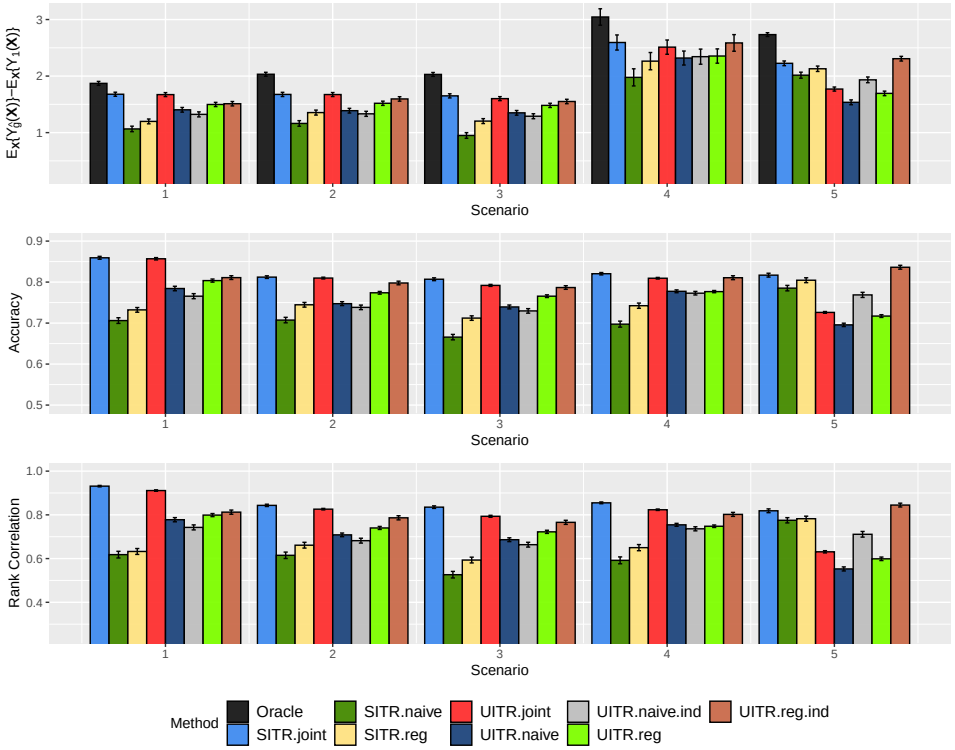

Fig 2: Average performance measures across the simulation replications are displayed. Confidence intervals are based on the standard errors of the results across the replications multiplied by 1.96.

"SITR.naive" and "SITR.reg" are more sensitive to the larger p as expected. This demonstrates the vulnerability of "separate" or group-wise analysis in the situation of small sample sizes and/or large number of covariates, and also indicates the benefit of joint analysis.

Comparing the results from Scenarios 3 and 4 which differ only in terms of the coefficient magnitudes from Study 1 (with stronger signal), we note that both the concordance rates and ranking correlations increase for "SITR.joint", "UITR.joint", "SITR.naive", "UITR.naive", "UITR.reg.ind", and "SITR.reg" but this increase is more modest for "UITR.naive" and "UITR.reg". This counter-intuitive phenomenon can be understood by further inspecting the results for each group displayed in Table 3 and Table 4 in the Supplementary Material. It is interesting to see that for "UITR.naive", "UITR.reg", and "UITR.reg.ind", although their performances in group 1 improve due to the

stronger signal in Scenario 4, performances for other groups are mild. On the other hand, "SITR.joint" and "UITR.joint" are able to improve their performances in all groups and have a more balanced performance across groups due to the usage of the standardization weight W^j in (2.11). This observation reveals potential issues with pooled analyses when treatment effects are highly variable across groups. Another case where pooled analyses might fail is Scenario 5. When groups are sufficiently different from each other and sample sizes are unbalanced, all the methods with a universal treatment rule tend to bias towards the dominant (i.e. with the largest sample size) group and typically perform poorly in the other groups. Note that in this case, "UITR.joint" seems to be a more robust choice than "UITR.naive" and "UITR.reg" and "UITR.reg.ind" performs the best, as, while it is fit in a pooled fashion, it allows for separate rules for each group due to the group by covariate interactions. Our investigations of the impact of main effect signal strength in the Supplementary Material suggest that the optimal choice of $a(\boldsymbol{X})$ for methods that use augmentation can mitigate the deterioration of performance that results from main effects with larger signal strength. Overall, the results for binary presented in the Supplementary Material tell a fairly similar story as the simulations with continuous outcomes. For binary outcomes, SITR.joint tends to perform nearly the best or the best across most settings, however for binary outcomes there is slightly more variability in which method works for specific studies. Interestingly, while "UITR.reg.ind" works well for continuous outcomes, it performs worst among all methods for binary outcomes.

5. Analysis of a TC program. In this section, we demonstrate the utility of our proposed framework via analysis of a TC intervention. We also compare with the different ITR estimation approaches from Section 4 by repeatedly randomly splitting the data into training and testing portions, fitting models on the training portions, and evaluating resulting ITRs on the outcomes of the testing portions. Following this, we also analyze the entire dataset using our proposed method.

5.1. *Problem formulation and modeling details.* Our aim is to evaluate the effect of a TC program on whether or not patients had a rehospitalization within 30-days of an index hospitalization (coded as 0, 1). Thus, the 30-day rehospitalization indicator is used as the outcome. Our analysis focuses on estimating ITRs that minimize the risk of 30-day rehospitalizations for patients. After excluding covariates that had little to no variability, we used 301 covariates in our analysis. They included various baseline covariates of patients such as gender and race and medical measurements such as anxiety

and hypertension. The analysis data set had 3869 medically uncomplicated subjects and 1007 medically complicated subjects.

We evaluated all approaches mentioned in Section 4. As the outcome for the TC analysis is binary, for "SITR.reg" and "UITR.reg" we utilized penalized logistic regression models to model the 30-day rehospitalization indicator. It is important to note that our proposed approach does not depend on the distribution of the outcome and can thus handle both binary and continuous outcomes without any changes to the underlying loss function.

The propensity scores were constructed using a penalized logistic regression model. The minimax concave penalty (MCP) was used as regularization term as it yielded the best marginal covariate balance in comparison with the lasso and SCAD penalties. The MCP-penalized logistic regression model was fit using the Orthogonalizing EM (OEM) algorithm of Xiong et al. (2016) computed with the `oem` package (Huling and Chien, 2018) due to the superior performance of the OEM algorithm for non-convex penalties compared with coordinate-descent algorithms. Cross validation using a cross-validated loss (2.8) with the penalty h set to 0 was utilized to select the penalization tuning parameter. The distributions of the propensity scores are plotted in Figure 3, which displays a reasonable overlap of the propensity scores between TC patients and controls. The balance of standard patient characteristics after inverse weighting by the resulting propensity scores is dramatically improved and is shown in Table 1 on the right hand side. The augmentation part of the AIPWE was based on a linear model for the outcome and the same AIPWE contrast estimator was utilized for all contrast classification based methods.

5.2. *Evaluation based on training and testing splits.* To evaluate the performance of the different methods in terms of their ability to yield ITRs that result in improved patient outcomes, we repeatedly randomly split the data into a training portion (3/4) and a testing portion (1/4), fitting models using the training portions and evaluating the impact of the ITRs on the corresponding testing portions. However, as the ground truth is never known in practice and it is thus impossible to know how a particular approach compares with the truly optimal ITRs, we could not evaluate the same performance measures as used in our simulation studies in Section 4. We therefore conducted evaluation by estimating the expected potential outcomes conditional on treatment recommendations.

Specifically, for any estimated ITR $\hat{g}$, we evaluated different methods using the following statistic,

$$\bar{Y}_{a,b}^{(\hat{g})} = \frac{\sum_{i=1}^{n} Y_i \mathbb{1}(A_i = a, \hat{g}(\boldsymbol{X}_i) = b)/P(A_i = a|\boldsymbol{X}_i)}{\sum_{i=1}^{n} \mathbb{1}(A_i = a, \hat{g}(\boldsymbol{X}_i) = b)/P(A_i = a|\boldsymbol{X}_i)} \tag{5.1}$$

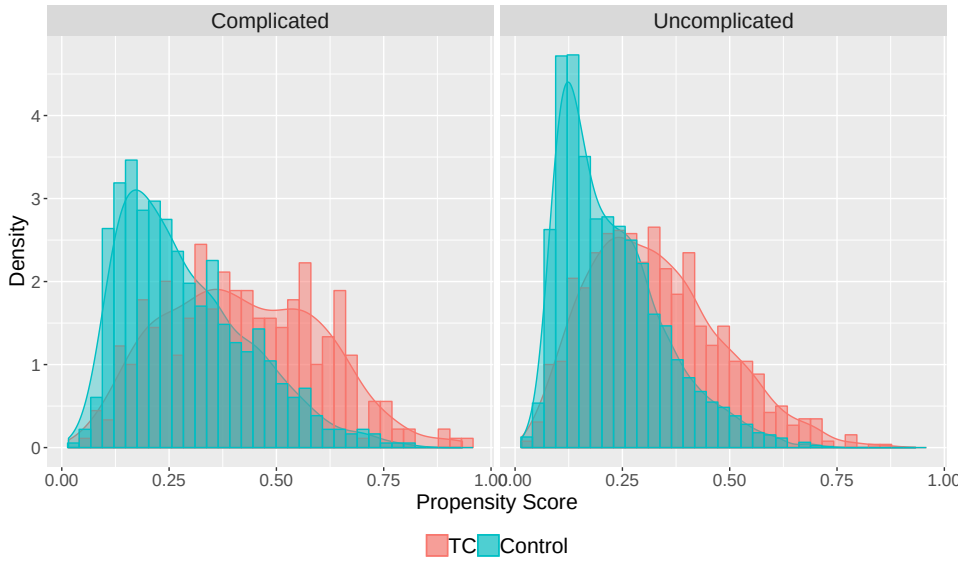

Fig 3: Displayed are the propensity score distributions for medically complicated and medically uncomplicated groups stratified by program enrollment status. The distributions have reasonable overlap across the range of possible values.

The following fact, which is proved in the Supplementary Material, justifies the usage of (5.1) to evaluate performance and relates it to potential outcomes.

Under the usual assumptions of the Rubin Causal Model (see Section 2.1),

$$\bar{Y}_{a,b}^{(\hat{g})} \xrightarrow{d} E(Y^{(a)}|\hat{g}(\boldsymbol{X}_i) = b) \tag{5.2}$$

for all $a, b \in \{0, 1\}$ as $n \to \infty$. Therefore if the treatment is indeed beneficial for those who are recommended to treatment by $\hat{g}$, we should expect a large value of $\bar{Y}_{1,1}^{(\hat{g})} - \bar{Y}_{0,1}^{(\hat{g})}$, which estimates $E(Y^{(1)}|\hat{g}(\boldsymbol{X}) = 1) - E(Y^{(0)}|\hat{g}(\boldsymbol{X}) = 1)$. Similarly, for the group of patients who are recommended to the control by $\hat{g}$, the expected improvement from changing treatment to control can be estimated as $\bar{Y}_{0,0}^{(\hat{g})} - \bar{Y}_{1,0}^{(\hat{g})}$, which estimates $E(Y^{(0)}|\hat{g}(\boldsymbol{X}) = 0) - E(Y^{(1)}|\hat{g}(\boldsymbol{X}) = 0)$. Alternatively, we expect the statistics of the concordant patients, $\bar{Y}_{0,0}^{(\hat{g})}$ and $\bar{Y}_{1,1}^{(\hat{g})}$ to be larger than those of the discordant patients, $\bar{Y}_{0,1}^{(\hat{g})}$ and $\bar{Y}_{1,0}^{(\hat{g})}$ for a good ITR $\hat{g}$.

Using (5.1), we evaluated all approaches across 1000 random splits of the dataset. Figure 4 displays these estimates averaged over 1000 test sets. The

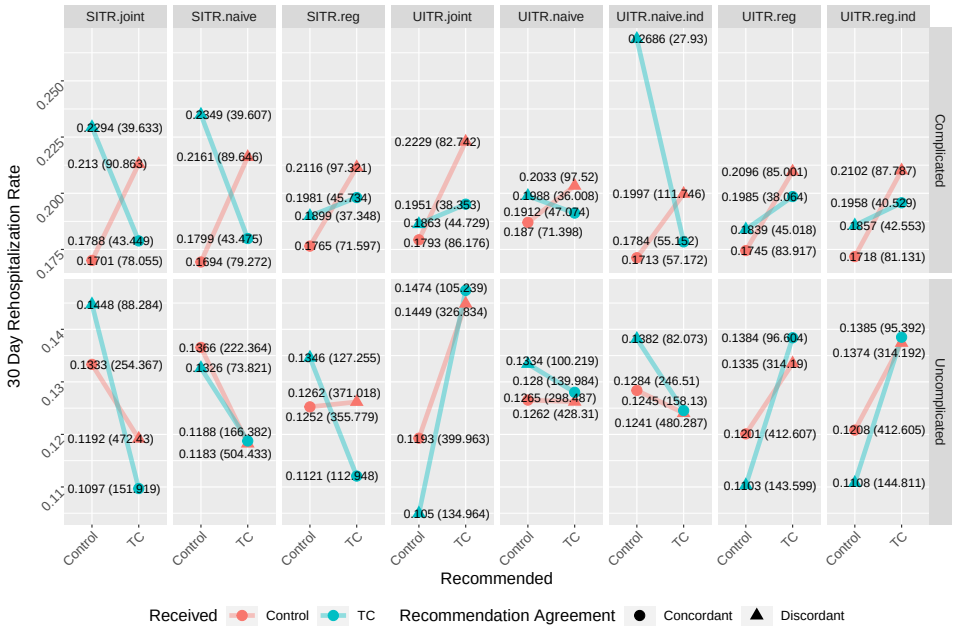

Fig 4: Training/test splits: interaction plots of 30 day rehospitalization rates for medically complicated and medically uncomplicated patients evaluated on the test datasets. The training and testing procedure was replicated 1000 times and results are averaged over the replications. Points labeled as circles are the average test set outcomes for scenarios when the intervention received is equal to the intervention recommendation and labeled as triangles otherwise.

corresponding average sample sizes are displayed within the parentheses. Within a given recommendation group (recommended TC or control), the difference between the displayed values of (5.1) represents the estimated reduction in the 30-day rehospitalization rate. We can see that separate ITRs for the two patient groups are necessary based on the patterns in these plots. That is, we generally see improved readmissions rates for those concordant patients than those discordant patients and the improvements are better when using separate ITRs. In particular, for the joint analysis and outcome regression, the universal ITR rule obviously leads to unsatisfactory results for the uncomplicated group. However the "SITR.naive" analysis also leads to unsatisfactory results for the uncomplicated group. The proposed approach "SITR.joint" has the best performance in terms of the reduction in 30-day rehospitalization rate for the medically uncomplicated group and performs

virtually the best for the medically complicated group.

5.3. *Results using the entire data.* We fit our proposed model using the entire training dataset. A total of 23 variables were selected into the estimated treatment rule for the complicated group, 40 were selected into the treatment rule for the uncomplicated group, 22 variables were selected in both the treatment rules for complicated and uncomplicated patients, 19 of which were estimated to have the same sign, and 19 variables were selected in just one of the two groups. In comparison, for the separate modeling approach (SITR.naive), 7 variables for selected into the ITR for medically uncomplicated patients and 3 were selected into the ITR for medically complicated patients and none of these were selected in common for both groups. All 10 of these variables were selected by SITR.joint. For the combined contrast analysis (UITR.naive), 13 variables were selected into the ITR, 12 of which were selected by SITR.joint. The cross validation error as a function of the tuning parameters is displayed in Figured 5. The minimizer of the cross validation error was $(\lambda, \tau_0) = (0.048, 0.398)$. Approximately 42% (423/1007) of medically complicated patients were recommended to TC, among whom 30% (127/423) received TC. The rest 58% (584/1007) of medically complicated patients were recommended usual care, among whom 65% (378/584) received the usual care. Approximately 63% (2440/3869) of medically uncomplicated patients were recommended TC, among whom 24% (583/2440) received TC. The rest 37% (1429/3869) of medically uncomplicated patients were recommended usual care, among whom 73% (1049/1429) received the usual care.

We now summarize several of the most impactful variables selected in the ITR by SITR.joint. The following characteristics tend to increase benefit of TC for both the medically complicated and uncomplicated groups: have lymph node swelling; nephritis, nephrosis, or renal sclerosis; those with fluid and electrolyte disorders; those with immune disorders; gastrointestinal disorders; those who had a claim with a provider whose specialty is medical oncology; those with personal exposures and history presenting hazards to health; and those who were prescribed hematological agents. Those (both medically complicated and uncomplicated) with symptoms involving nervous and musculoskeletal systems tend to benefit from TC less. Medically uncomplicated patients with paralysis are more likely to benefit, while medically complicated patients are less likely to benefit, medically uncomplicated patients with radiologic guidance were more likely to benefit, medically complicated patients with immune disorders were more likely to benefit, while medically uncomplicated patients with immune disorders were less likely to

benefit.

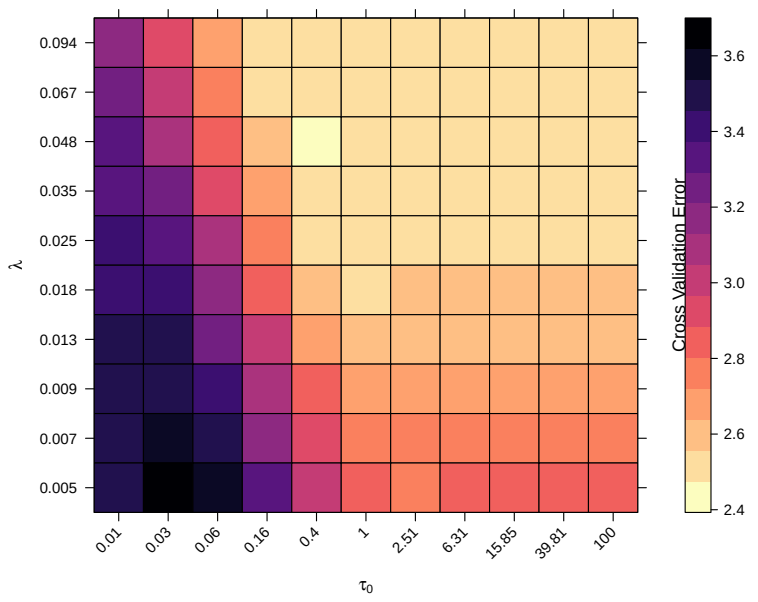

Fig 5: Displayed are the cross validation errors averaged across the folds versus a grid of values for λ and τ_0 for the proposed method fit using the entire data.

6. Conclusion. Motivated by the example of TC, we have proposed an ITR recommendation framework suitable for settings where there are different patient groups that behave quite differently from each other. By incorporating group structure of coefficients into a variable selection procedures, it can be more efficient than methods which analyze each patient group separately. The proposed approach also automatically adjusts for the heterogeneity in the magnitudes of treatment effects across the different groups, which is a common issue in practice, especially in scenarios where the treatment may feasibly be implemented slightly differently for different groups. We also propose a similar framework which provides a universal treatment rule for all the patient groups. This may be useful when the group membership of a new patient is not immediately known, e.g. when the DRG of a patient is not coded on admission.

The proposed subgroup identification framework is not limited to the problem of multiple patient groups and can easily be adapted to many other

settings. For example, individual patient data (IPD) meta-analysis, which aims to achieve a higher statistical power and more robust point estimates through aggregation of information from multiple studies, could be fit into the proposed framework by treating each study as a patient group. In this case, the term (2.10) may be more helpful if the studies included in a meta-analysis are more similar to each other. Consequently, the ITRs are also more likely to be close to each other in terms of coefficient magnitudes.

As another example, the proposed framework is also applicable to settings where multiple outcomes are measured for each subject. To accommodate this problem using the proposed framework, we can duplicate each subject by the number of outcomes observed and treat one copy of subjects and one outcome as a patient "group". Specifically, using the notation of the proposed framework, we let Y_i^j denote the jth outcome of the ith subject, and $\boldsymbol{X}_i^j$, A_i^j as the jth copy of baseline covariates and the treatment of the ith subject. In this manner, the setting of modeling multiple outcomes can be viewed as a special case of our setting, where $n_1 = \cdots = n_q$, $\boldsymbol{X}_i^1 = \cdots = \boldsymbol{X}_i^q$ and $A_i^1 = \cdots = A_i^q$; our framework is then readily applicable. In multiple outcomes, the choice of W_j in (2.11) is essential, as different outcomes are very likely to have different scales, especially when they are not of the same data type.

There are a few possible extensions to make it even more flexible and robust. For instance, same as in Zhang et al. (2012), our framework use the squared loss as a surrogate for the 0-1 loss and estimate the treatment rule from (2.3). The squared loss, nevertheless, is vulnerable to outliers. A more robust surrogate loss function could be employed and the consistency in Theorem 2.1 should be maintained. However, the specific form of $a(\boldsymbol{X})$ in Proposition 2.2 must be re-derived, since it depends on the particular loss function used. Another possible extension is to move beyond linear classifiers and utilize a more flexible class of decision rules. This may bring added challenges, as even with an additive model it is not immediately apparent how best to borrow strength in variable selection of the additive functions.

Acknowledgments. Research reported in this article was partially funded through two Patient-Centered Outcomes Research Institute (PCORI) Awards (ME-1409-21219 and HSD-1603-35039). The views in this publication are solely the responsibility of the authors and do not necessarily represent the views of the PCORI, its Board of Governors or Methodology Committee. This project was also supported by the UW Health Office of Population Health, the Health Innovation Program, the UW School of Medicine and Public Health from The Wisconsin Partnership Program, and the Community-

Academic Partnerships core of the University of Wisconsin Institute for Clinical and Translational Research (UW ICTR) through the National Center for Advancing Translational Sciences (NCATS), grant UL1TR000427. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Supplementary Materials. The proofs of all theoretical results, additional simulation results, and codes are available in the Supplementary Materials.

References.

Betancourt, J. R., Tan-McGrory, A. and Kenst, K. (2015). Guide to preventing readmissions among racially and ethnically diverse Medicare beneficiaries. *Health*.

Bradley, E. H., Curry, L., Horwitz, L. I., Sipsma, H., Thompson, J. W., Elma, M., Walsh, M. N. and Krumholz, H. M. (2012). Contemporary evidence about hospital strategies for reducing 30-day readmissions: a national study. *Journal of the American College of Cardiology* **60** 607–614.

Bradley, E. H., Curry, L., Horwitz, L. I., Sipsma, H., Wang, Y., Walsh, M. N., Goldmann, D., White, N., Piña, I. L. and Krumholz, H. M. (2013). Hospital strategies associated with 30-day readmission rates for patients with heart failure. *Circulation: Cardiovascular Quality and Outcomes* **6** 444–450.

Chen, S., Tian, L., Cai, T. and Yu, M. (2017). A general statistical framework for subgroup identification and comparative treatment scoring. *Biometrics* **73** 1199–1209.

Cloonan, P., Wood, J. and Riley, J. B. (2013). Reducing 30-day readmissions: Health literacy strategies. *Journal of Nursing Administration* **43** 382–387.

Coleman, E. A., Parry, C., Chalmers, S. and Min, S.-j. (2006). The Care Transitions Intervention: Results of a Randomized Controlled Trial. *Archives of Internal Medicine* **166** 1822-1828.

Cox, D. R. (1958). *Planning of Experiments*. Wiley.

Donzé, J., Aujesky, D., Williams, D. and Schnipper, J. L. (2013). Potentially avoidable 30-day hospital readmissions in medical patients: derivation and validation of a prediction model. *JAMA Internal Medicine* **173** 632–638.

Evashwick, C. (2005). *The Continuum of Long-term Care*. Cengage Learning.

Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association* **96** 1348–1360.

Fox, T., Brummit, P. S., Ferguson-Wolf, M., Abernethy, M. et al. (2000). Position of the American Dietetic Association: Nutrition, aging, and the continuum of care. *Journal of the Academy of Nutrition and Dietetics* **100** 580.

Hansen, L. O., Young, R. S., Hinami, K., Leung, A. and Williams, M. V. (2011). Interventions to reduce 30-day rehospitalization: a systematic review. *Annals of Internal Medicine* **155** 520–528.

Holland, P. W. (1986). Statistics and causal inference (with discussion). *Journal of the American Statistical Association* **81** 945–960.

Huling, J. D. and Chien, P. (2018). Fast Penalized Regression and Cross Validation for Tall Data with the oem Package. *Journal of Statistical Software*. To appear.

Imai, K. and Ratkovic, M. (2013). Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation. *Annals of Applied Statistics* In press.

Jencks, S. F., Williams, M. V. and Coleman, E. A. (2009). Rehospitalizations among patients in the Medicare fee-for-service program. *New England Journal of Medicine* **360** 1418–1428.

Kehl, V. and Ulm, K. (2006). Responder identification in clinical trials with censored data. *Computational Statistics & Data Analysis* **50** 1338–1355.

Kind, A. J., Brenny-Fitzpatrick, M., Leahy-Gross, K., Mirr, J., Chapman, E., Frey, B. and Houlahan, B. (2016). Harnessing Protocolized Adaptation in Dissemination: Successful Implementation and Sustainment of the Veterans Affairs Coordinated-Transitional Care Program in a Non-Veterans Affairs Hospital. *Journal of the American Geriatrics Society* **64** 409–416.

Kripalani, S., Theobald, C. N., Anctil, B. and Vasilevskis, E. E. (2014). Reducing hospital readmission rates: Current strategies and future directions. *Annual Review of Medicine* **65** 471–485.

Leppin, A. L., Gionfriddo, M. R., Kessler, M., Brito, J. P., Mair, F. S., Gallacher, K., Wang, Z., Erwin, P. J., Sylvester, T., Boehmer, K. et al. (2014). Preventing 30-day hospital readmissions: a systematic review and meta-analysis of randomized trials. *JAMA Internal Medicine* **174** 1095–1107.

Lipkovich, I., Dmitrienko, A. and B D'Agostino Sr, R. (2017). Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine* **36** 136–196.

McIlvennan, C. K., Eapen, Z. J. and Allen, L. A. (2015). Hospital readmissions reduction program. *Circulation* **131** 1796–1803.

Naylor, M. D., Aiken, L. H., Kurtzman, E. T., Olds, D. M. and Hirschman, K. B. (2011). The importance of transitional care in achieving health reform. *Health Affairs* **30** 746–754.

Norrving, B. and Kissela, B. (2013). The global burden of stroke and need for a continuum of care. *Neurology* **80** S5–S12.

Ollier, E. and Viallon, V. (2017). Regression modelling on stratified data with the lasso. *Biometrika* **104** 83–96.

Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *Annals of Statistics* **39** 1180.

Rau, J. (2014). Medicare fines 2,610 hospitals in third round of readmission penalties. *Kaiser Health News* **2**.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of Regression Coefficients When Some Regressors are not Always Observed. *Journal of the American Statistical Association* **89** 846-866.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association* **100** 322–331.

Shi, C., Song, R., Lu, W. and Fu, B. (2018). Maximin projection learning for optimal treatment decision with heterogeneous individualized treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2013a). A Sparse-Group Lasso. *Journal of Computational and Graphical Statistics* **22** 231-245.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2013b). SGL: Fit a GLM (or cox model) with a combination of lasso and group lasso regularization R package version 1.1.

Stevens, S. (2015). Preventing 30-day readmissions. *Nursing Clinics* **50** 123–137.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B* 267–288.

Xiong, S., Dai, B., Huling, J. and Qian, P. Z. G. (2016). Orthogonalizing EM: A

design-based least squares algorithm. *Technometrics* **58** 285–293.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* **38** 894–942.

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M. and Laber, E. (2012). Estimating optimal treatment regimes from a classification perspective. *Stat* **1** 103–114.

Zhao, Y., Zeng, D., Rush, A. and Kosorok, M. (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association* **107** 1106–1118.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B* **67** 301–320.

Menggang Yu
Department of Biostatistics & Medical Informatics
Health Innovation Program
University of Wisconsin-Madison
Madison, WI 53706
E-mail: meyu@biostat.wisc.edu

Chensheng Kuang
Department of Statistics
University of Wisconsin-Madison
Madison, WI 53706
E-mail: chenshengkuang@gmail.com

Jared D. Huling
Division of Biostatistics
School of Public Health
University of Minnesota
Minneapolis, MN 55455
E-mail: huling@umn.edu

Maureen Smith
Department of Population Health Sciences,
Department of Family Medicine & Community Health, and
Health Innovation Program
University of Wisconsin-Madison
Madison, WI 53706
E-mail: maureensmith@wisc.edu